Journal of Biomedical Informatics 131 (2022) 104110

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

The IeDEA harmonist data toolkit: A data quality and data sharing solution for a global HIV research consortium

Judith T. Lewis [a,b,*], Jeremy Stephens [c], Beverly Musick [d], Steven Brown [d], Karen Malateste [e], Cam Ha Dao Ostinelli [f], Nicola Maxwell [g], Karu Jayathilake [h], Qiuhu Shi [i], Ellen Brazier [j,k], Azar Kariminia [l], Brenna Hogan [m], Stephany N. Duda [a,n], on the behalf of leDEA

[a] Vanderbilt Institute for Clinical and Translational Research, Vanderbilt University Medical Center, Nashville, TN, USA
[b] Department of Biomedical Engineering, Vanderbilt University, Nashville, TN, USA
[c] Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA
[d] School of Medicine, Indiana University, Indianapolis, IN, USA
[e] French National Research Institute for Sustainable Development (IRD), Inserm, UMR 1219, University of Bordeaux, Bordeaux, France
[f] Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland
[g] Centre for Infectious Disease Epidemiology and Research, School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa
[h] Department of Infectious Diseases, Vanderbilt University Medical Center, Nashville, TN, USA
[i] Department of Public Health, New York Medical College, Valhalla, NY, USA
[j] Institute for Implementation Science in Population Health, City University of New York, New York, NY, USA
[k] Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA
[l] The Kirby Institute, UNSW Sydney, Australia
[m] Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[n] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA

ARTICLE INFO

Keywords:
Data harmonization
HIV
Global health
Biomedical informatics
Data quality

ABSTRACT

We describe the design, implementation, and impact of a data harmonization, data quality checking, and dynamic report generation application in an international observational HIV research network. The IeDEA Harmonist Data Toolkit is a web-based application written in the open source programming language R, employs the R/Shiny and RMarkdown packages, and leverages the REDCap data collection platform for data model definition and user authentication. The Toolkit performs data quality checks on uploaded datasets, checks for conformance with the network's common data model, displays the results both interactively and in downloadable reports, and stores approved datasets in secure cloud storage for retrieval by the requesting investigator. Including stakeholders and users in the design process was key to the successful adoption of the application. A survey of regional data managers as well as initial usage metrics indicate that the Toolkit saves time and results in improved data quality, with a 61% mean reduction in the number of error records in a dataset. The generalized application design allows the Toolkit to be easily adapted to other research networks.

1. Introduction

The International epidemiology Databases to Evaluate AIDS (IeDEA) consortium is an international research network formed in 2006 by the U.S. National Institute of Allergy and Infectious Diseases (NIAID) to allow investigators to address high priority HIV/AIDS research questions through the combination and analysis of globally diverse clinical data [1]. Merging observational data from participating sites in a research network like IeDEA presents challenges for data harmonization and data quality; assurance of data quality is critical to the integrity of any analysis [2–6]. Like many disease-focused research consortia, IeDEA is a federated network with limited resources for technical support, so data quality solutions that require local software installation and dedicated data quality personnel are generally not feasible. Therefore, a need exists for a straightforward, user-friendly software solution for data quality checking, reporting, and secure data sharing in such research consortia. We describe the development and impact of the Harmonist Data Toolkit, our approach to streamlining data harmonization and

* Corresponding author at: Vanderbilt Institute for Clinical and Translational Research, 2525 West End Ave Suite 1050, Nashville, TN 37203-8820, USA.
E-mail address: judy.lewis@vumc.org (J.T. Lewis).

https://doi.org/10.1016/j.jbi.2022.104110
Received 2 August 2021; Received in revised form 4 February 2022; Accepted 1 June 2022
Available online 6 June 2022
1532-0464/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

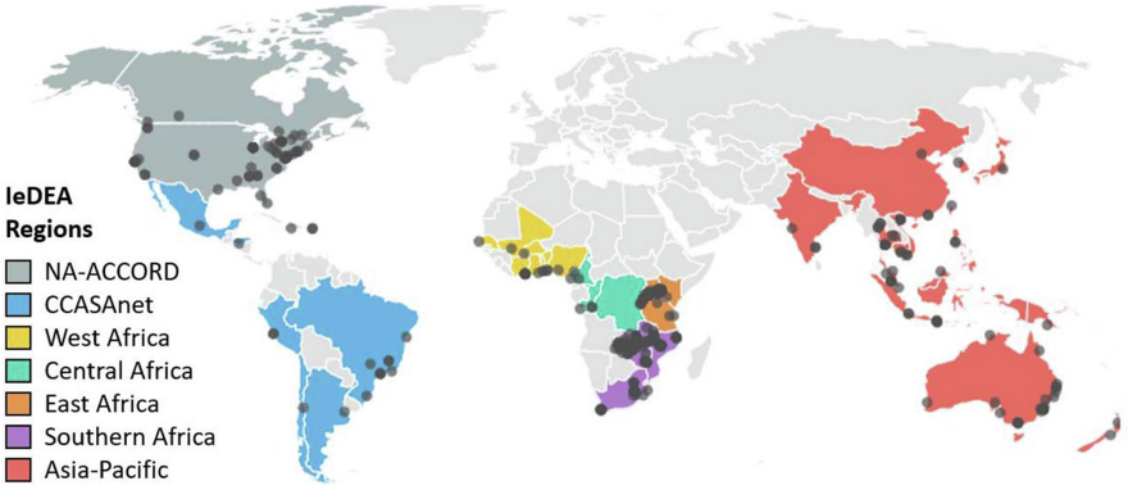

Fig. 1. IeDEA regions and participating sites.

improving data quality in IeDEA by leveraging the Research Electronic Data Capture (REDCap) platform [7].

1.1. IeDEA

IeDEA is composed of seven regional networks that collect data from over 380 HIV clinics worldwide, representing over 2 million adults and children living with and at risk for HIV [1] (Fig. 1). Clinics in IeDEA are grouped into these seven geographical regions, each with its own regional data center that is staffed by 1–4 data managers. These data managers are responsible for collecting and harmonizing observational data—in a wide variety of formats and languages—from the sites in their region. IeDEA regions collaborate by combining data from their regional databases for use in approved multiregional analyses [8–12]. Due to data privacy concerns and the data sharing regulations of participating countries and institutions, IeDEA regions contribute subsets of the data collected from clinics in their region for approved projects rather than pooling data in a centralized database. The U.S. National Institute of Allergy and Infectious Diseases purposefully developed this federalized structure for IeDEA to build a foundation of trust among global collaborators with differently resourced countries. Once a multiregional research study is approved by IeDEA regions, each participating region contributes data. Regional data managers select observations that match the study's inclusion and exclusion criteria and submit a dataset that has been mapped to the IeDEA Data Exchange Standard (DES), a common data model for observational HIV data [13]. Historically, the work of mapping data to comply with the latest version of the data model and performing extensive data quality checking required substantial data manager effort. As a result, the research process could be slow and individual study data submissions often could not be readily merged into a global dataset by the data recipient. A more efficient and practical solution to real-time data mergers was needed to leverage the global data resources within IeDEA.

1.2. Data quality approaches in other research networks

Other research collaborations have addressed data quality assessment with a variety of methods. In a landmark 2016 paper [4], Kahn et al examined data quality approaches and terminology in over twenty large data-sharing networks and found significant inconsistencies in the ways that quality issues were evaluated, described, and reported. To improve transparency and understanding of data quality, they proposed a standardized framework of intrinsic data quality concepts based on three categories: Conformance (do data values match the constraints of the common data model?), Completeness (are data values present?), and Plausibility (are data values believable?). In 2017, Callahan et al [5] expanded upon this work by mapping six research networks' data quality checks to the categories established by Kahn, summarizing differences in the coverage of data quality checks, and investigating the types of data quality tools and personnel resources used. They found that all six networks employed complex processes that included installing and updating required software—either open source or proprietary—and depended on the availability of both local and centralized data quality support, as well as familiarity with technical tools like GitHub. Liaw et al [14] conducted an extensive literature review of data quality practices in real world data and saw the need for an expanded data quality framework, one that includes meaningful but hard to measure contextual and technical indicators such as timeliness, trustworthiness, and traceability.

These studies reveal that technically demanding data quality control practices are used by both centralized and distributed data networks. For example, data quality management is centralized at a coordinating center in the National Patient-Centered Clinical Research Network (PCORnet) [15–17], Pediatric Learning Health System (PEDSnet) [5], and the Sentinel Initiative [18–20]. These three networks use the Observational Medical Outcomes Partnership (OMOP) common data model, and query their network sites about data quality using software that is installed and run at sites locally. Data quality issues are reviewed by coordinating center personnel who follow up with the sites. Data quality control is maintained centrally. In contrast, the Observational Health Data Science and Informatics (OHDSI) network has distributed data quality coordination [21–23]. The original data quality tool developed for OHDSI, Achilles Heel [21,24], functioned primarily as a dataset characterization tool and was recently supplemented by the OHDSI Data Quality Dashboard [25]. The Data Quality Dashboard applies 20 "data quality ideas" to OMOP-formatted datasets: 8 conformance checks, 5 completeness rules, and 7 checks for data plausibility. Depending on the number of tables and fields in a dataset, this can add up to as many as 3,300 data quality checks.

These existing data preparation and data quality tools, although robust and well-suited to their networks, require high-level technical expertise and server infrastructure at either the participating sites or a central coordinating center. As a network with no coordinating center, a majority of sites located in resource-limited settings, and limited availability of IT support and infrastructure at the regional level, IeDEA needed an adaptable solution. In this paper, we present the collaborative design method and initial use results for the Harmonist Data Toolkit, an intuitive web-based data quality checking, report generation, and data exchange application designed for IeDEA but generalizable to other research domains with data models defined in REDCap.

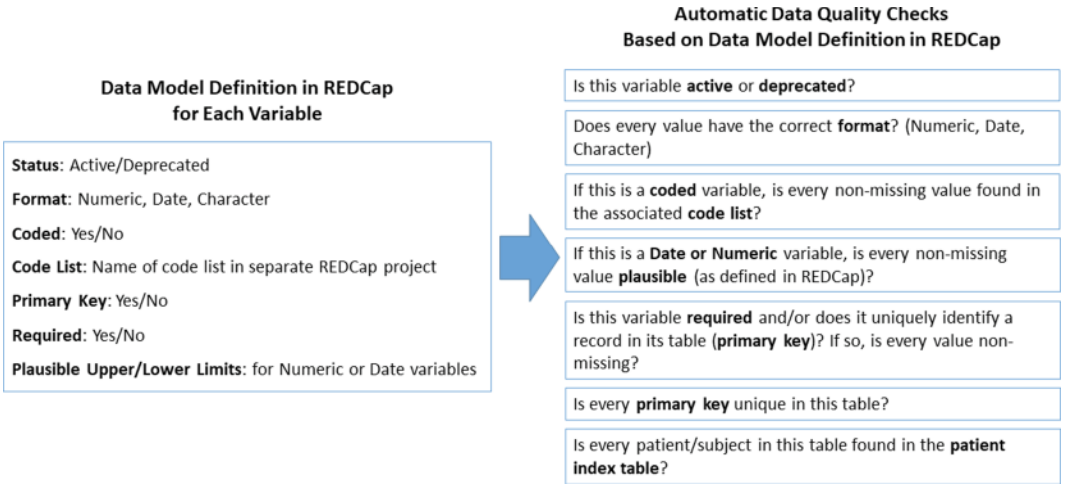
Data Model Definition in REDCap for Each Variable
Status: Active/Deprecated
Format: Numeric, Date, Character
Coded: Yes/No
Code List: Name of code list in separate REDCap project
Primary Key: Yes/No
Required: Yes/No
Plausible Upper/Lower Limits: for Numeric or Date variables
Automatic Data Quality Checks Based on Data Model Definition in REDCap
Is this variable active or deprecated?
Does every value have the correct format? (Numeric, Date, Character)
If this is a coded variable, is every non-missing value found in the associated code list?
If this is a Date or Numeric variable, is every non-missing value plausible (as defined in REDCap)?
Is this variable required and/or does it uniquely identify a record in its table (primary key)? If so, is every value non-missing?
Is every primary key unique in this table?
Is every patient/subject in this table found in the patient index table?

Fig. 2. Abstracting data model details in REDCap. Data quality checks based on these details automatically include new variables and codes.

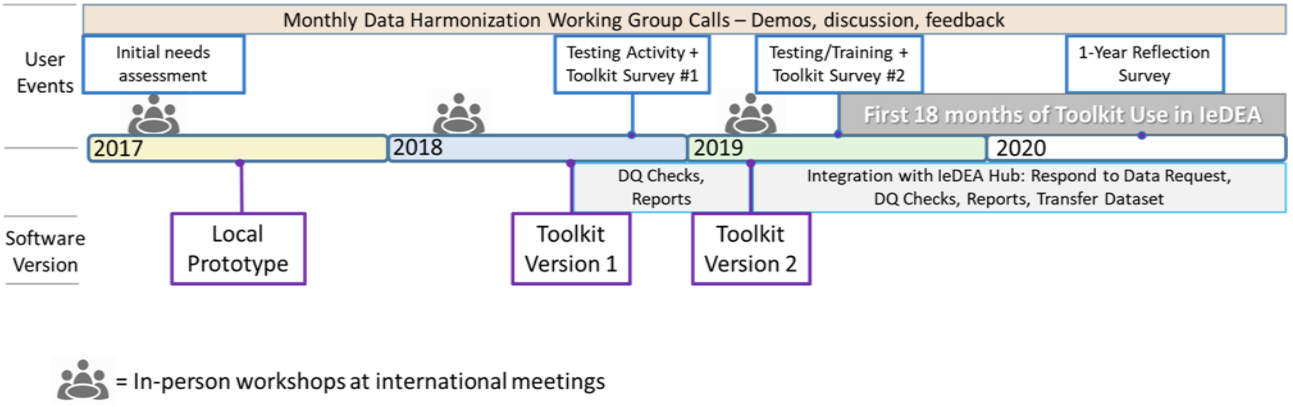
Monthly Data Harmonization Working Group Calls – Demos, discussion, feedback
User Events
Initial needs assessment
Testing Activity + Toolkit Survey #1
Testing/Training + Toolkit Survey #2
1-Year Reflection Survey
First 18 months of Toolkit Use in IeDEA
2017
2018
2019
2020
DQ Checks, Reports
Integration with IeDEA Hub: Respond to Data Request, DQ Checks, Reports, Transfer Dataset
Software Version
Local Prototype
Toolkit Version 1
Toolkit Version 2
= In-person workshops at international meetings

Fig. 3. Collaborative design timeline for IeDEA Harmonist Data Toolkit.

2. Methods

2.1. Engaging stakeholders and identifying software needs

Throughout the software development process, our informatics team sought both qualitative and quantitative input from IeDEA data managers and investigators through in-person and virtual meetings, email discussions, one-to-one calls, and REDCap surveys. We participated in monthly calls with IeDEA data managers (the IeDEA Data Harmonization Working Group) and hosted side meetings at international research conferences to collaborate in person with the community of stakeholders. Due to the small number of intended users of the IeDEA Harmonist Data Toolkit (1–4 data managers at each of the seven regional centers), it was not necessary to employ random sampling when distributing surveys or choosing participants for meetings and calls; we engaged data representatives from every region in the needs assessment stage as well as the software design, testing, and revision processes. Similarly, with such a small group of stakeholders, there was no need for formal analysis of meeting notes and responses to open-ended survey questions. Each REDCap user survey was introduced during a working group call, distributed to all data managers via email, and the responses were exported from REDCap into R [26] for tabulation. The first of these REDCap surveys, an initial Needs Assessment Survey, sought feedback from every data manager about their current workflow, challenges they face in data quality and data exchange, and their software preferences.

2.2. Designing data quality checks for IeDEA

Members of the IeDEA Data Harmonization Working Group determined the scope of desired data quality checks based on their experience with IeDEA multiregional studies and common errors in IeDEA datasets. Those data quality checks were a combination of tests that individual data managers had implemented locally as well as checks used in other research networks [2,5,15,24,27]. The Harmonist team mapped the data quality checks to the data quality categories proposed by Kahn [4].

Next, we modeled the IeDEA common data model in a series of REDCap databases representing data tables, variables, and code lists. This allowed the data model to be easily edited and exported from REDCap in a machine-readable JSON format. We developed data checking scripts using the R statistical computing language [26] that processed the JSON data model as input and implemented general data quality checks based on the details from REDCap (Fig. 2). This approach allowed us to abstract—rather than hard-code—the data quality checks, and ensured applied checks would always be up-to-date with the current version of the data model and code lists in REDCap.

To keep data quality feedback useful and relevant, the agenda of the monthly Data Harmonization Working Group included regular discussions on new checks proposed by data managers as well as suggested modifications to current checks.

2.3. Implementing a user-centered, iterative software design approach

After developing an initial prototype of the Harmonist Data Toolkit,

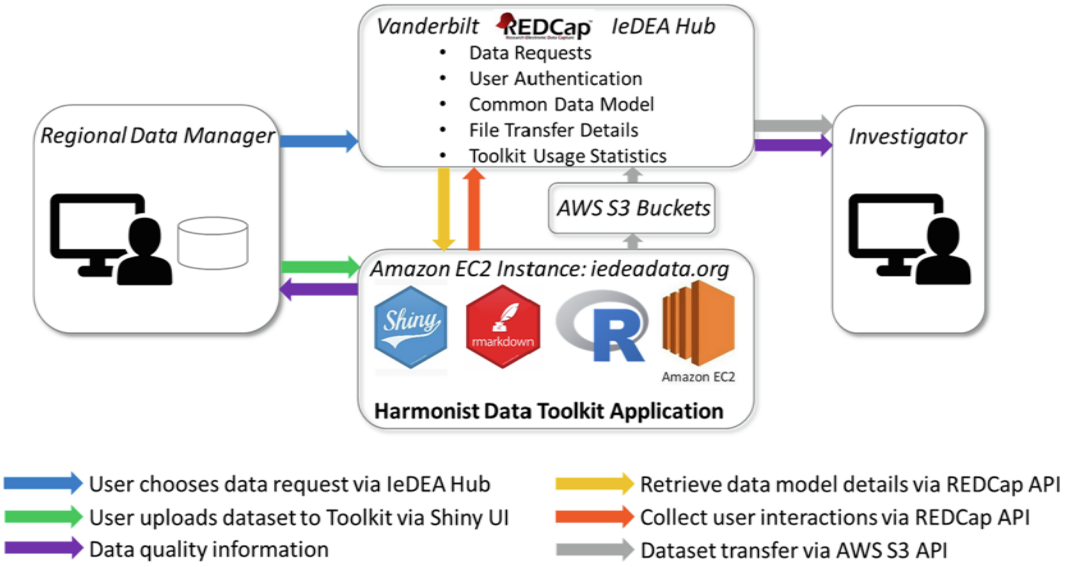

Fig. 4. Harmonist Data Toolkit system architecture and communication with REDCap.

we iterated toward our current application design and report content through several cycles of demonstrations, feedback gathering, and software modifications (Fig. 3). This included two separate asynchronous user testing events with assigned tasks to complete, followed by surveys. In the first testing activity, users were asked to explore the features of the Toolkit and complete the survey. Instructions for Testing/Training Activity #2 were more specific and included every activity that a data manager would need to perform, including uploading datasets, performing data quality checks, generating reports, submitting datasets, deleting datasets, and retrieving submitted datasets. These tasks were chosen to test the full functionality of the Toolkit so that users could both learn and evaluate every aspect of the new workflow. Sample datasets were provided so that users could assess their feelings about Toolkit security without using real data. Both training events were introduced in a Data Harmonization Working Group call followed by emails to all data managers with details, sample datasets, and links to follow-up surveys in REDCap. Toolkit Survey #1 included a combination of open-ended and multiple-choice questions about which Toolkit features seemed most useful, how the Toolkit would impact the user's workflow, as well as what changes they would suggest making to the software and report content. In Toolkit Survey #2, users were asked to rate their comfort with using the Toolkit for IeDEA data tasks and to list any remaining questions or steps of the process they found confusing. To encourage candid responses, this survey was anonymous.

2.4. System architecture

We implemented the Harmonist Data Toolkit as a web application written in the open source R statistical programming language [26]. It leverages the Shiny [28] web framework for the user interface, the RMarkdown [29] package for reporting, and other R packages for support including ggplot2 [30] and tidyverse [31]. The Toolkit application interfaces with REDCap via the REDCap API to retrieve details about the common data model and data requests as well as for Toolkit usage tracking (Fig. 4).

The application is deployed on an Amazon EC2 instance and is accessible via the web for data quality checking and report generation for IeDEA DES-compliant datasets. The Toolkit can also be used in conjunction with IeDEA multiregional research data requests for users who access the Toolkit by way of a REDCap-based consortium management portal, known as the IeDEA Hub, which requires user authentication. Data managers who log in to the IeDEA Hub are able to transfer datasets in response to data requests. Submitted datasets are stored in Amazon S3 buckets, to be retrieved by specified data downloaders after multi-factor authentication. All data transactions, including downloads, uploads, and deletions, are logged in the IeDEA Hub. Uploaded datasets are automatically deleted from Amazon S3 buckets after 30 days. Our Amazon Web Services contract includes a Business Associate Addendum to ensure that protected health information is safeguarded. IeDEA regional data centers ensure compliance with all data sharing and data de-identification regulations. The Harmonist team operated under established institutional data use agreements as the Harmonist project was developed by investigators in the IeDEA Latin America region.. One region requested an additional investigator-signed data agreement to ensure sample datasets they provided for testing would be used only for testing.

2.5. Measuring Toolkit usage and impact

2.5.1. Surveying user experience

After the first year of active Toolkit use, we asked for user feedback via the anonymous 1-Year Reflection REDCap survey. We distributed the survey link by email to all IeDEA data managers and asked them to evaluate their experiences with the Toolkit in comparison to previous methods of data quality checking, data sharing, and project management. The results of this survey were exported from REDCap into R for analysis.

2.5.2. Quantifying Toolkit usage and dataset errors

In order to ascertain Toolkit acceptance by users, learn which features of the Toolkit were used most often, and track the number and types of errors in user datasets over time, we created a REDCap project that creates a new record for specific actions performed by Toolkit users. Each record documents the type of action chosen by the user and the timestamp of the action. For example, when a user uploads a dataset, the names of the tables and variables are stored in the REDCap record. When data quality checks have been completed, the REDCap record includes a summary of the types of errors found in the dataset. For users who accessed the Toolkit directly by visiting the URL, the records are anonymous and are not linked to a user or region. When users have entered the Toolkit by logging in to the IeDEA Data Hub and selecting a data request, the tracking records identify the user, their region, and the data request number.

After tracking Toolkit activity through December 2020, we exported

Table 1
Software requests reported by IeDEA regional data managers.

REQUEST CATEGORY 1: ENSURE EASE OF USE
Avoid software that requires local installation or maintenance
Minimize need for technical resources or personnel
Enable users to work in desired data environments (SAS, Stata, CSV/Excel, R, SPSS)
Provide intuitive, easy-to-use interface
Generate automatic, customizable reports
REQUEST CATEGORY 2: SIMPLIFY DATA HARMONIZATION
Flag variables, formats, and codes that are not consistent with the data model
Check for plausible numeric values and dates, date order logic, and agreement between related variables (e.g., an end date and a reason for ending)
Offer data quality results interactively and in downloadable spreadsheets
Provide an error spreadsheet that includes all details necessary to locate records with data quality issues
Keep data quality checks up to date with the current version of the common data model
REQUEST CATEGORY 3: PROVIDE WORKFLOW SUPPORT
Remind users of details of tables and variables requested for a selected study
Prioritize dataset security and privacy (no long-term data storage)
Transfer files submitted for multiregional studies using a secure approach
Generate dataset summaries and data quality reports automatically when datasets are transferred, to provide a history of dataset submissions

Table 2
Example Toolkit data quality checks based on data model details in REDCap, mapped to the harmonized data quality assessment categories proposed by Kahn and colleagues [4].

General Data Quality Rule	Example Data Quality Check	Data Quality Category
Primary key values must be non-null	*PATIENT, VIS_D (visit date), and CENTER must be complete in tblVIS*	Completeness
Each patient identifier in every table must link to a record in the patient index table	*Patients in tblART (antiretroviral therapy data table) must be listed in tblBAS (patient index table)*	Conformance: Relational
Primary keys must be unique	*The combination of PATIENT, ART_ID (ART medication code), and ART_SD (ART medication start date) must be unique for each record in tblART*	Plausibility: Uniqueness
Variables flagged as required in the DES must have non-null values	*Each patient in tblBAS should have a non-missing BIRTH_D value*	Completeness
First dates in date pairs should precede second dates	*MED_SD (medication start date) should be before MED_ED (end date)*	Plausibility: Temporal
Values for coded variables should be found in the associated code list	*Valid values for DROP_Y (Has patient dropped out?) are 0 (No) or 1 (Yes)*	Conformance: Value
Numeric values should be plausible	*Patient weight (kg) should be less than 200 and greater than 0.5*	Plausibility: Atemporal

records from REDCap into R and analyzed the first 18 months of Toolkit use. We then wrote scripts in the R programming language to summarize and visualize user interactions and data quality results over time. Software testing and development sessions were excluded from this analysis.

2.5.3. Determining Toolkit impact on dataset quality

In many cases, users uploaded and revised datasets multiple times for a single data request before ultimately submitting the files. This provided an opportunity to compare errors before and after a user viewed Toolkit data quality results, since the inclusion/exclusion criteria, tables, and variables were constant. To observe changes in data quality after repeated sessions with the Toolkit, we selected only those sets of multiple uploads by a single user for a specific data request. We tracked the number and types of errors in each upload/data quality check cycle for each of these groups of uploads, from the initial to the final data quality check before dataset transfer.

Using the error summaries on these sets of datasets that were uploaded and revised, we wrote an additional R script to determine which types of errors occur most frequently in datasets and which types of errors are most commonly corrected. Among the datasets with multiple uploads and revisions, the script compared the number of initial and final datasets that contained each type of error.

3. Results

3.1. Software requirements/constraints

In their responses to the initial Needs Assessment survey, IeDEA data managers who *prepare and send* datasets for IeDEA multiregional studies (n = 9) identified multiple challenges that they encounter when responding to IeDEA multiregional data requests: mapping datasets to the IeDEA common data model; cleaning data; finding a secure way to send datasets, confirm receipt, and later track what was sent; responding to questions from the investigator or data manager who received the dataset; and managing resubmissions of datasets due to feedback from the requesting investigator. IeDEA data managers who also *receive and merge* datasets for IeDEA multiregional studies (n = 6) reported the following challenges in their survey responses: receiving datasets with nonstandard variable names and codes; datasets with missing data; late data submissions; implausible dates in datasets; and the lack of a consistent, secure method of receiving datasets from other regions.

We reviewed these survey results along with our notes from meetings with IeDEA data managers and compiled a list of user priorities to guide our design of an IeDEA data quality software solution (Table 1).

3.2. Data quality checks in Toolkit

The data quality checks are conducted on each dataset, table, and variable, according to the details stored in REDCap, as described in Section 2.2. As soon as a new table or variable is added to the data model and defined in REDCap, any uploaded dataset that includes that table or variable will automatically be checked for duplicate records, correct variable format, completeness, valid codes if applicable, plausible values for numeric and date variables, and correct date order logic. Similarly, when a new code is added to the data model, it is automatically included as a permissible value for all variables associated in REDCap with that code list.

When a data quality issue is detected in the dataset, a new entry is added to the Toolkit's list of dataset errors. Each entry in the list includes the table name(s) and primary key values needed for a data manager to locate the record(s) related to the data quality issue. Each error is tagged with a severity level: Critical, Error, or Warning. Critical errors are those that present significant problems for dataset analysis and integrity, such as missing or duplicate patient IDs in tables with patient ID as the sole primary key. Examples of IeDEA data quality checks and corresponding errors are shown in Table 2. (See the Appendix for further descriptions of Harmonist data quality checks.).

Additional data quality information is presented in the form of visualizations in the Toolkit reports. Data managers and investigators can assess temporal plausibility and completeness issues by reviewing histograms of the number of observations by date (Fig. 5a). Similarly, heat maps reflecting the percentage of patients from each site/patient group who are included in each table can highlight gaps in data collection (Fig. 5b). These visualizations were added to the Toolkit report based on requests from IeDEA investigators who were interested in detecting incomplete data from individual sites.

3.3. Redcap Testing/Training survey results

In November 2018, following the release of Version 1 of the Toolkit, the nine primary IeDEA regional data managers representing all seven IeDEA regions were invited to complete a series of assigned tasks and submit responses to Toolkit Survey #1. All nine data managers

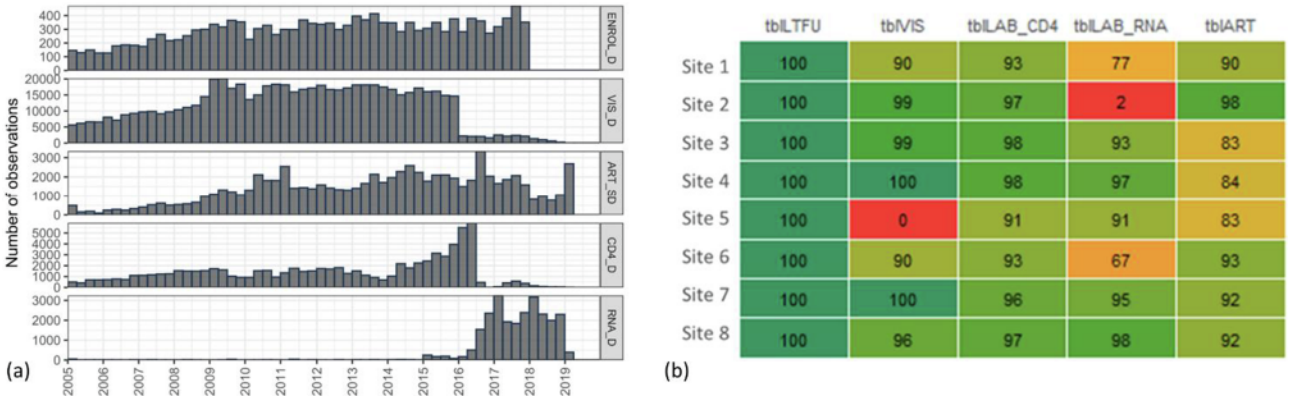

Fig. 5. Report visualizations useful in data quality assessment: (a) Example of histograms of enrollments, clinic visits, lab tests, ART medication initiation, and disease diagnoses by date for each site. Investigators can spot unusual trends, such as the drop off in documented clinic visits after 2015 for this example site. (b) Heat maps of patient representation in data tables (e.g., loss to follow-up from clinic [LTFU], visits, CD4 cell count lab results, HIV viral load lab results, and antiretroviral therapy [ART]) draw attention to gaps in reporting, such as the lack of any clinic visit data from "Site 5" in the example above.

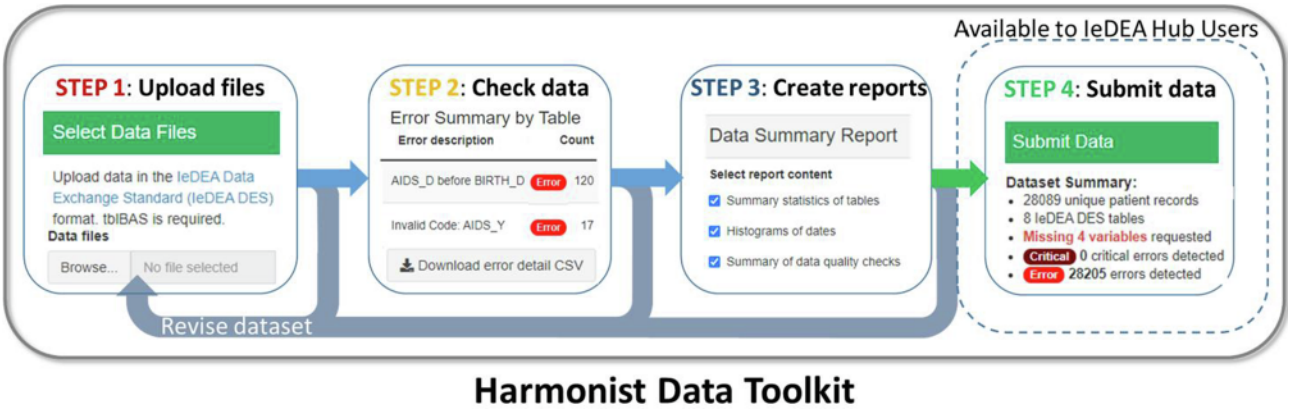

Fig. 6. Harmonist Data Toolkit workflow overview.

participated and identified themselves in the survey. The ability to download error detail spreadsheets was ranked as the most useful Toolkit feature, followed by interactive review of data quality results and automatically generated reports. In response to open-ended questions about the Toolkit, users suggested additional features, including customizing report names and enabling users to download separate error spreadsheets for each patient group and/or data table. These features were incorporated into Version 2 of the Toolkit. All respondents indicated that they envisioned the Toolkit would save time and improve data quality.

To introduce Version 2 of the Toolkit, test its behavior, and encourage its adoption, we conducted a comprehensive Testing/Training Activity followed by Toolkit Survey #2 in June 2019. Seventeen data-related personnel from all IeDEA regional data centers were emailed step-by-step instructions, a sample dataset, and a REDCap survey link. All 17 participated in the training activity and 14 of the 17 (82%) completed the anonymous feedback survey. When asked to report their comfort on a scale of 0–100 (100 = "Completely comfortable") with using the Toolkit to upload, check, and share datasets for IeDEA multiregional studies, the mean response was 92 (min = 75, max = 100, n = 14). In addition, 93% (13) of respondents answered "No" to the question: "Were any of the [Toolkit] steps confusing?" In this REDCap survey, we also asked users to submit questions that remained unanswered after the exercise and to list features of the Toolkit they would like to understand better. In response, we added a FAQ page to the Toolkit that addressed those questions.

3.4. Application workflow

We listened to feedback from users throughout the design process and modified the user interface and functionality as requested. This included implementing the Harmonist Data Toolkit as a web-based application, eliminating the need for local installation or maintenance of software.

The resulting application consists of 4 main steps (Fig. 6): (1) *Upload files*, in the user's preferred data format, (2) *Check data*, using the common data model definition in REDCap as an up-to-date guide, (3) *Create reports*, to download and share, and (4) *Submit data* to the investigator requesting the data. Step 4 is available to users who have logged in to the IeDEA Hub and selected an active data request prior to Step 1.

Step 1: Upload data.

The Toolkit prompts users to browse and select the file(s) containing the dataset. If a user has accessed the Toolkit through the IeDEA Hub in response to a multiregional data request, the Toolkit reminds the user which tables are requested for this study. A link to the complete data request is available for more details about study inclusion and exclusion criteria.

Data managers may upload a single ZIP file or select multiple individual files, whose names must match the table names of the IeDEA Data Exchange Standard. The files can be in the user's preferred format (SAS, Stata, CSV, or SPSS). At a minimum, the dataset must include the patient index table with one record per patient and each table in the dataset must include the primary key variables. Each patient ID in the patient

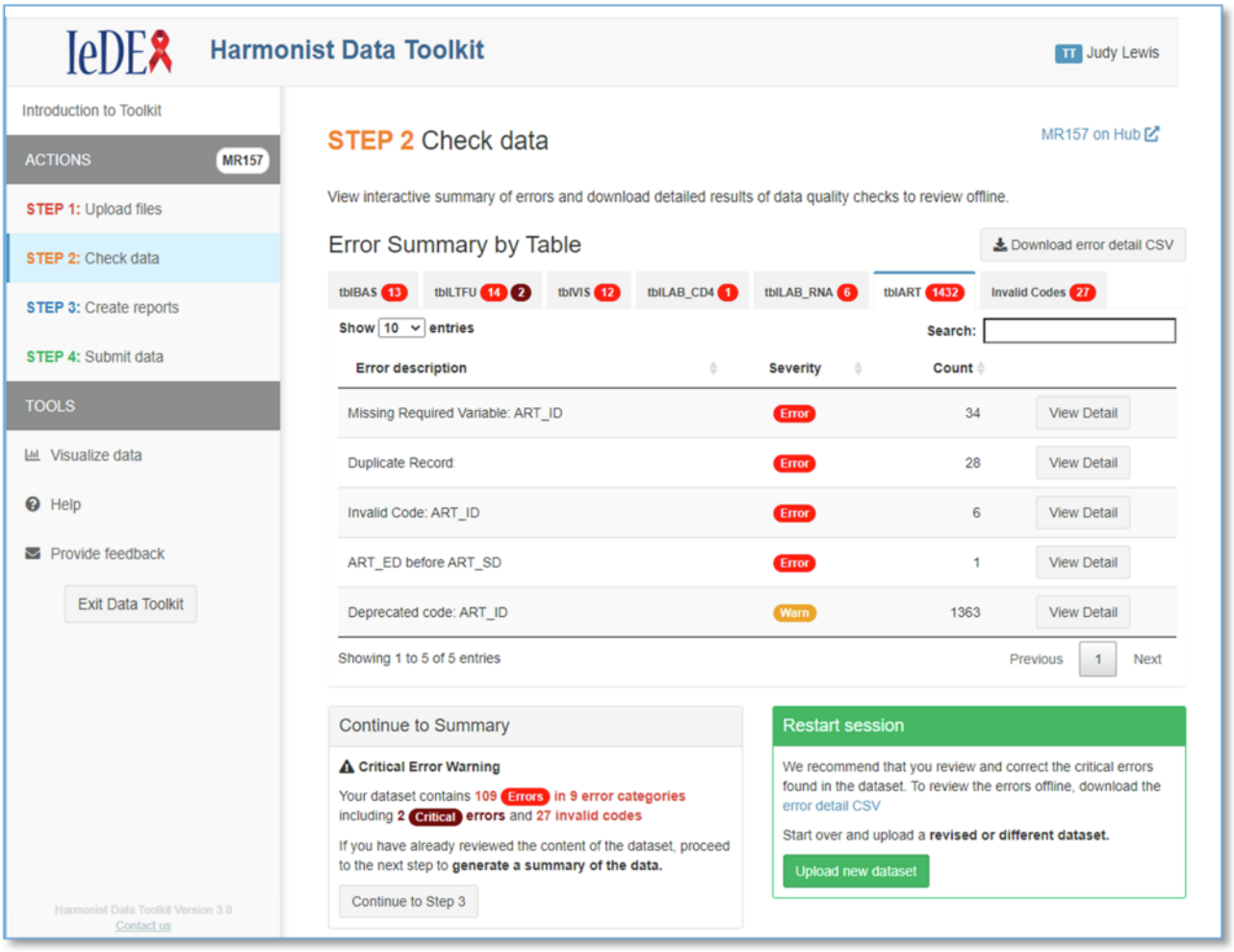

Fig. 7. Screenshot of Step 2 of the Harmonist Data Toolkit.

index table must be unique.

If these conditions are met, the user can review and confirm the list of IeDEA tables and variables that were detected in the selected files. Non-IeDEA files may be uploaded, such as text or PDF files that convey important information about the dataset to the investigator. These files will be ignored by the data quality checks but will be available to authenticated designated data recipients if the dataset is ultimately submitted to secure cloud storage for a multiregional data request. Before users continue to Step 2, they are notified of any missing requested tables and variables (if logged in to a specific data request) and encouraged to revise and upload again.

Step 2: Check data.

As the dataset is checked for conformance, plausibility, and completeness, the application displays the progress through the data quality checks. Once all data quality checks are complete, a summary of the results for each IeDEA table are displayed in an interactive webpage (Fig. 7). A badge beside each table name indicates the number of errors found in that table. The summary tables report the number of errors for each variable and error type. The summary tables are sortable by error description, severity, or prevalence. To see details that identify the specific records containing that error, users may click the "View Detail" button in that row.

As requested by data managers, we included a "Download error detail CSV" option so that users can track down each erroneous record offline and correct it if possible. Users can choose to divide the error details into multiple spreadsheets by table and/or patient group or clinic. This option streamlines the process of locating and correcting sources of error at the regional level.

Step 3: Create summary.

The Toolkit provides two customizable report options: an overall report and a data quality metrics report. The overall report includes descriptive statistics of the dataset, a summary of data quality checks, histograms of important dates, and heat maps summarizing data completeness (Fig. 5). Users can tailor report content and choose to include all patients, a single patient group, or create individual reports for each patient group. The quality metrics report, created in response to feedback from IeDEA investigators, displays the completeness and quality of data for each variable within a table, for each site included the dataset.

Step 4: Submit data.

For authenticated users who access the Toolkit through the IeDEA Hub in response to a specific data request, an additional step is available. This final step allows users to transfer the uploaded dataset to secure cloud storage. Clicking the "Submit Data" button zips the files shared by the user, and communicates via the AWS API to store the dataset in a new secured AWS S3 bucket. In addition, the "Submit Data" process triggers generation of a standard overall report to be stored in a REDCap Data Uploads project along with the name of the corresponding S3 bucket. The report is available to authorized data downloaders on the IeDEA Hub. Datasets are automatically deleted after 30 days.

Optional step: Dataset visualization.

We added a visualization option to allow users to explore their dataset and download publication-ready plots.

Consistent user interface design.

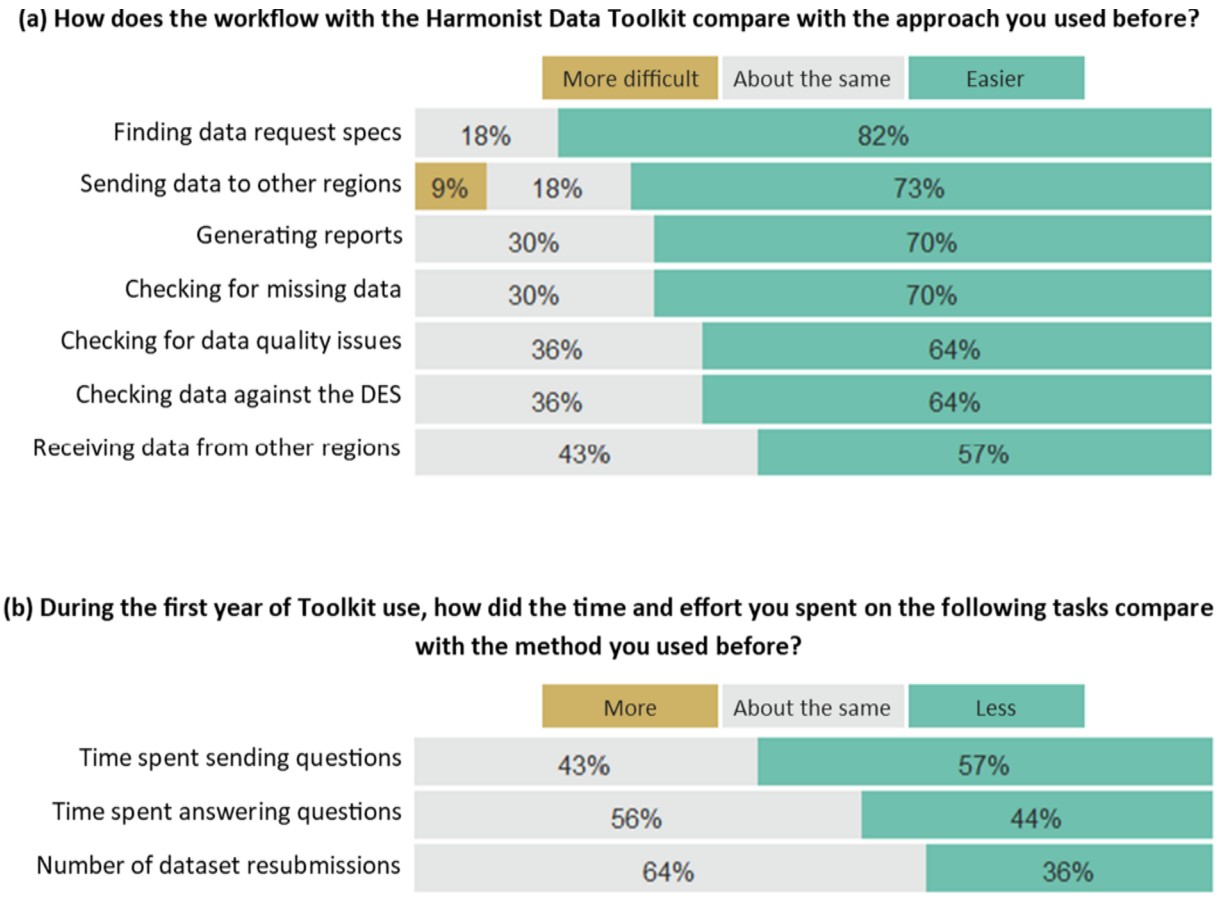

Fig. 8. Results of REDCap Survey of IeDEA data managers after the first year of Toolkit use in IeDEA. Data managers compared Toolkit workflow with their previous methods of data quality checking and data sharing for IeDEA multiregional studies.

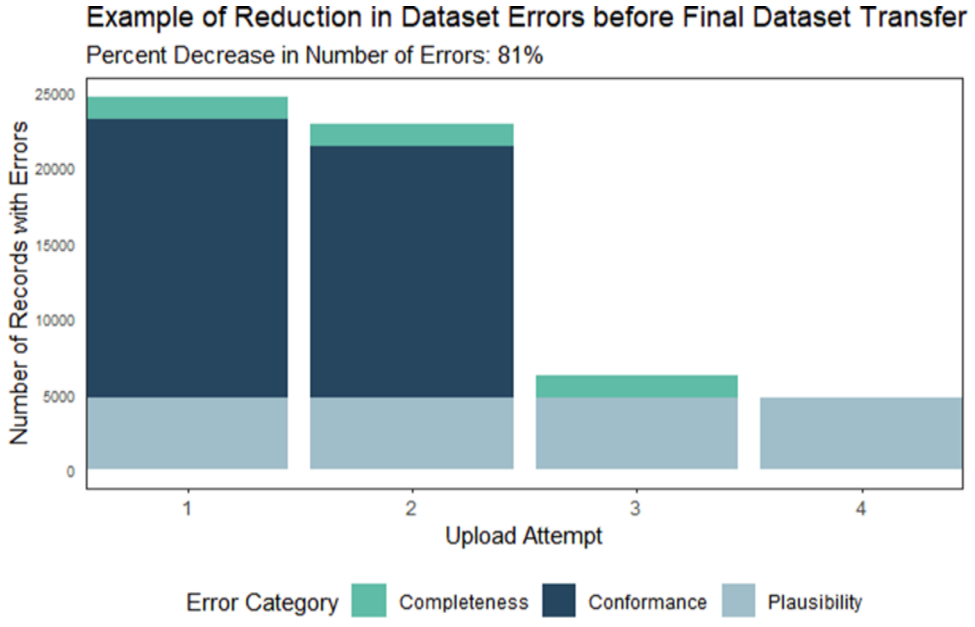

Fig. 9. Example analysis of dataset errors in a series of uploads and revisions by a single data manager for a specific data request. On the final iteration, the dataset was transferred to the investigator who requested the data. See the Appendix for additional graphs.

Throughout the application, we maintained a cohesive user interface with consistent color choices. We also followed a convention of presenting the user with two boxes at the bottom of each page: a green box highlighting the recommended next step and a gray box providing an alternative option. These colors are chosen dynamically depending on the severity of errors found in the dataset.

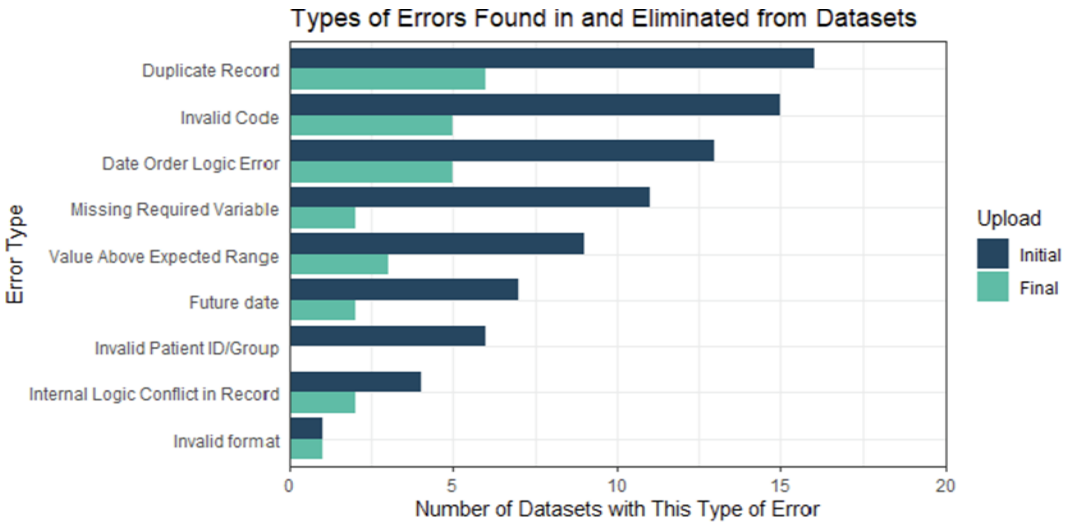

Fig. 10. Error types found in initial uploads as compared with final uploads among datasets that were checked with the Toolkit multiple times and revised before submission.

Tutorials and help page.

Users who enter the Toolkit directly through the URL land on a welcome page which explains the workflow and includes links to the IeDEA data model browser, the IeDEA Hub, a sample data to use in testing the Toolkit, and a YouTube video tutorial. The Toolkit also includes a Help page with an exhaustive Frequently Asked Questions section, informed by questions submitted by data managers in the Toolkit Testing REDCap Survey. These questions and answers are retrieved from a REDCap project which allows for easy updates to the FAQ section.

3.5. Toolkit use and impact on data quality

3.5.1. User Toolkit evaluation after one year of use

In July 2020, one year after initial adoption of the Toolkit, we invited 14 regional data managers representing all seven IeDEA regions to respond to a 1-Year Reflection Survey in REDCap. Of these, 13 (93%) completed the anonymous survey and shared their experiences with the Toolkit, including their perceptions of its impact on workflow and data quality.

Survey respondent roles within IeDEA included the following: prepare datasets to submit for IeDEA multiregional projects (n = 13), receive or merge datasets for multiregional projects (n = 9), and analyze datasets (n = 7). All respondents had used the Toolkit for the following activities: testing and training activities (n = 11), checking and correcting data (n = 11), submitting datasets for multiregional projects (n = 11), and generating reports (n = 9).

Survey participants indicated that Toolkit use has made IeDEA data management tasks easier (Fig. 8). Notably, 100% of respondents reported that Toolkit use had improved the quality of IeDEA datasets. When asked about the Toolkit's automatically generated reports, all users responded that the reports were useful in multiple ways, including checking for data gaps and revising datasets. Among the 11 data managers who used the Toolkit to submit datasets for multiregional requests, 91% (n = 10) reported having revised datasets based on Toolkit data quality reports before transferring data to investigators.

3.5.2. Toolkit adoption and usage in first 18 months

Analysis of the first 18 months of Toolkit usage as tracked in REDCap revealed that the Toolkit was used by data managers from all seven IeDEA regions to check, summarize, and transfer datasets that included patients from every clinic in their region. Datasets were uploaded to the Toolkit 507 times for quality checking, revealing a total of over 269 million data errors and warnings across these datasets. Approximately 35% of data quality check sessions were initiated through the IeDEA Hub in response to an IeDEA multiregional data request, whereas 65% were anonymous sessions initiated by accessing the URL directly (without access to dataset transfer). The number of patients in a single dataset ranged from 457 to 986,089, with a mean of 123,769 patients per dataset. A total of 260 dataset reports were downloaded by users, not including the reports automatically generated and stored on the Hub to document dataset transfer. Of all uploads, 41 final datasets and linked reports were transferred to secure cloud storage through the Toolkit in response to multiregional data requests. Use of the Toolkit for dataset transfer varied across the seven regional data centers: One region transferred 13 datasets through the Toolkit, two regions transferred 7 datasets each, three regions transferred 4 datasets each, and one region transferred 2 datasets during this 18-month window of time.

3.5.3. Impact of Toolkit use on data quality

Of the 41 datasets that were shared for multiregional studies in the first 18 months of Toolkit use, 20 were uploaded, checked, and revised before final dataset transfer. In all cases, the datasets were transferred to the requesting investigator with fewer errors than were present in the initial upload. In 6 cases, 100% of errors were corrected. The mean percent decrease in the number of errors between the first data quality session with the Toolkit and the final dataset submission was 61% (max = 100%, min = 0.003%). Fig. 9 shows an example of the decrease in the number of dataset records that triggered errors with each cycle of checking a dataset and analyzing data quality feedback. (See the Appendix for error reduction visualizations for all twenty datasets.).

Among these 20 datasets, the most common type of error found in the initial upload was the existence of duplicate records (n = 16 datasets). After Toolkit data quality feedback and subsequent dataset revision, 6 datasets contained duplicate records. Invalid codes for coded variables (e.g., medication code) were found in 15 of the 20 datasets initially, but were present in 5 of the final, transferred datasets. The variable most commonly coded incorrectly was ART_ID, the ATC code for patient antiretroviral medication. Fig. 10 summarizes how many datasets included errors of each type on initial upload and which types of errors users were able to correct after receiving data quality feedback from the Toolkit.

4. Discussion

Incorporating the Harmonist Data Toolkit into the IeDEA data

workflow has improved the quality of datasets submitted for studies and increased transparency around data quality. The Toolkit has simplified the task of linking data quality checks to a data model and, unlike many other data quality solutions [5,15–24], it does not require locally installed and maintained software. Use of the Toolkit is popular among IeDEA data managers; they have observed new efficiencies from the streamlined workflow, automated data quality feedback, sharable reports, and the way that the Toolkit tracks the history of final dataset submissions. Improving data preparation and submission processes allows for more timely analyses. High data quality is critical for the results of these analyses to be trusted and have an impact on patient care and health policy.

4.1. Lessons learned

Frequent contact between the application development team and IeDEA data managers not only resulted in an application well-suited to users' needs but also increased user comfort with and adoption of the tool. A few data managers were initially slow to learn the new process, but usage tracking data showed that each time a testing and training activity or a demonstration was conducted, more data managers began using the Toolkit and the frequency of sessions increased. Adoption of the Toolkit by every region required working to gain the trust of the investigators and data managers in each region. By taking steps to ensure data privacy (allowing each region to review data requests and choose whether or not to participate, deleting datasets from Toolkit server memory as soon as a Toolkit session ends, limiting transferred dataset access to previously designated data downloaders, requiring multifactor authentication of data downloaders, logging all data uploads, downloads, and deletions in the IeDEA Data Hub, and automatically deleting datasets from cloud storage after 30 days), we demonstrated respect for the sensitive data the Toolkit processes and transfers.

4.2. Limitations

The architecture and implementation of the Toolkit has several limitations. The application relies on having a data model defined as tables, variables, code lists, and associated metadata in a REDCap template. Although this approach encompasses the data structures used by many research consortia, it may not map to highly abstracted data models that rely on semantic linkages from external ontologies. The data quality tests performed by the Toolkit application were defined by users, and likely do not cover the full spectrum of useful data quality checks. Networks interested in the Toolkit may need to add custom R data quality checks when adapting the open source code for their data model and use case. In particular, cross-variable data quality checks that depend on the semantics rather than the syntax of the data would require custom R scripts that are hooked into the Toolkit framework.

Although using a cloud service provider delivered a robust server infrastructure for our users, it may not be appropriate for all installations of the Toolkit. Experienced network and server managers are needed to set up a secure cloud environment. If using the Toolkit to process datasets subject to heightened data privacy regulations (e.g., European General Data Protection Regulation [GDPR]), additional planning and documentation with legal counsel is needed.

Our implementation used the free version of Shiny Server, which does not support multithreading. Since our users were distributed across time zones and the Toolkit processed datasets quickly, we were able to handle all requests sequentially. If parallel processes were needed, however, to conduct data quality checks in other networks with simultaneous user sessions involving large datasets, it would require additional servers or a paid license for Shiny Server Pro for multithreading support.

Finally, we recognize that the Toolkit is not an ideal solution for all types of research networks. Large research networks with established data quality tools and processes, such as PCORnet or OHDSI, would not benefit from adoption of the Toolkit. Similarly, networks processing large EHR-extracted datasets should use software designed for data in OMOP or HL7 Fast Healthcare Interoperability Resources (FHIR) formats [32]. The Toolkit fills the gap of available tools for less highly-resourced research networks facing data harmonization challenges who use disease-specific data models and have limited technical resources available for data quality.

4.3. Future work

Our next priority is to adapt the Toolkit for use beyond IeDEA while also continuing our active collaboration with the IeDEA Data Harmonization Working Group to improve the Toolkit's data quality assessments and reports. To facilitate use of the Toolkit by other research networks, we are currently in the process of generalizing the Toolkit code to remove IeDEA-dependent elements (e.g., logos, consortium details, variable naming conventions) and store those details in REDCap. The resulting application will accommodate any data model that has been defined in standardized REDCap templates. After testing the generalized Toolkit with a different data model and dataset, we plan to create documentation, including custom data quality check templates, for our public GitHub repository to guide other groups in Toolkit implementation. To expand data quality checks to include medication and other codes beyond those entered manually into the data model in REDCap, we intended to enhance support for use of standard terminologies by linking to the BioPortal API [33]. Other future work includes extending the coverage of Toolkit report content and data quality checks as suggested by Liaw et al [14] to include contextual data quality assessments, such as indicators of data recentness. We also plan to specify thresholds of acceptable levels of each type of error, a practice common in other research networks.

5. Conclusion

The importance of data quality in research networks is well established. For federated research networks similar to IeDEA, with no central database and limited availability of regional technical personnel and data resources, a web-based software solution for data quality checking, report generation, and dataset transfer like the Harmonist Data Toolkit can improve data quality and increase research throughput. Furthermore, by abstracting the data quality checking algorithms based on a common data model definition in REDCap, we designed data quality checks that automatically adapt to an evolving data model. Because software tools can only be effective if they are embraced by users, our collaboration with IeDEA data managers throughout the design process and implementation, coupled with structured testing and training activities, were key to the Toolkit's impact and success. Our application is open source [34] and our approach is suitable for generalization to other research networks with a data model defined in REDCap.

Contributors

Author contributions: JL and SD conceived the project, took notes during meetings, gathered input from stakeholders, and conducted training. All authors provided input on the design. JL and JS implemented the software. All authors tested the software. JL drafted the manuscript. All authors reviewed, edited, and approved the final manuscript.

Funding Statement

The Harmonist project is supported by the U.S. National Institutes of Health's National Institute of Allergy And Infectious Diseases, National Institute of Diabetes and Digestive and Kidney Disease, National Library of Medicine, and National Institute on Alcohol Abuse and Alcoholism

under Award Number R24AI124872. The International epidemiology Databases to Evaluate AIDS (IeDEA) is supported by the National Institute of Allergy and Infectious Diseases, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, the National Cancer Institute, the National Institute of Mental Health, the National Institute on Drug Abuse, the National Heart, Lung, and Blood Institute, the National Institute on Alcohol Abuse and Alcoholism, the National Institute of Diabetes and Digestive and Kidney Diseases, the Fogarty International Center, and the National Library of Medicine: Asia-Pacific, U01AI069907; CCASAnet, U01AI069923; Central Africa, U01AI096299; East Africa, U01AI069911; NA-ACCORD, U01AI069918; Southern Africa, U01AI069924; West Africa, U01AI069919. This work is solely the responsibility of the authors and does not necessarily represent the official views of any of the institutions mentioned above.

CRediT authorship contribution statement

Judith T. Lewis: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Jeremy Stephens:** Methodology, Software, Writing – review & editing. **Beverly Musick:** Methodology, Validation, Data curation, Writing – review & editing. **Steven Brown:** Validation, Data curation, Writing – review & editing. **Karen Malateste:** Validation, Data curation, Writing – review & editing. **Cam Ha Dao Ostinelli:** Validation, Data curation, Writing – review & editing. **Nicola Maxwell:** Validation, Data curation, Writing – review & editing. **Karu Jayathilake:** Validation, Data curation, Writing – review & editing. **Qiuhu Shi:** Validation, Data curation, Writing – review & editing. **Ellen Brazier:** Validation, Data curation, Writing – review & editing. **Azar Kariminia:** Validation, Data curation, Writing – review & editing. **Brenna Hogan:** Validation, Data curation, Writing – review & editing. **Stephany N. Duda:** Conceptualization, Methodology, Validation, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix

Fig. A1 and Table A1.

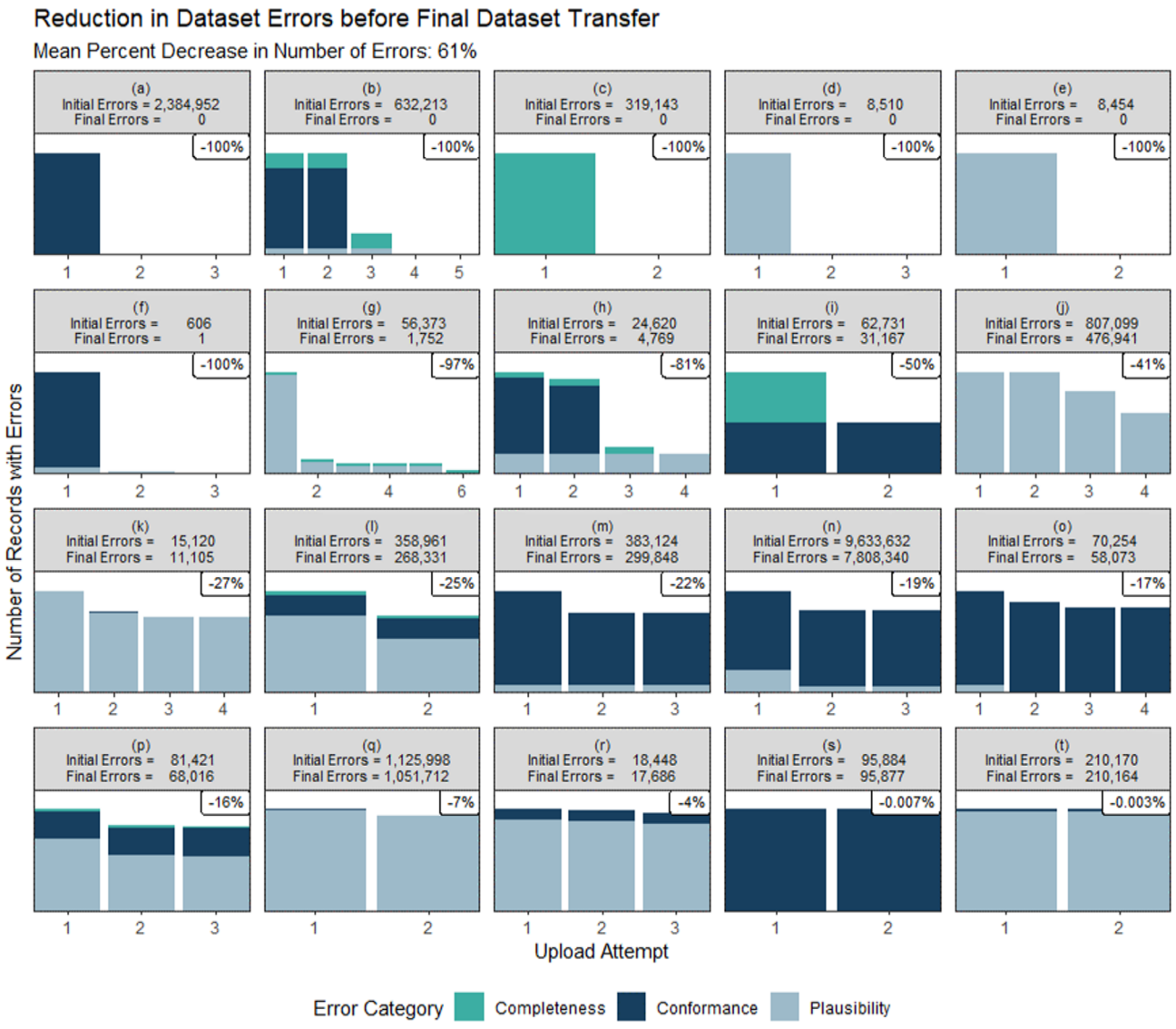
Reduction in Dataset Errors before Final Dataset Transfer
Mean Percent Decrease in Number of Errors: 61%
Number of Records with Errors
Upload Attempt
(a) Initial Errors = 2,384,952 Final Errors = 0 -100%
(b) Initial Errors = 632,213 Final Errors = 0 -100%
(c) Initial Errors = 319,143 Final Errors = 0 -100%
(d) Initial Errors = 8,510 Final Errors = 0 -100%
(e) Initial Errors = 8,454 Final Errors = 0 -100%
(f) Initial Errors = 606 Final Errors = 1 -100%
(g) Initial Errors = 56,373 Final Errors = 1,752 -97%
(h) Initial Errors = 24,620 Final Errors = 4,769 -81%
(i) Initial Errors = 62,731 Final Errors = 31,167 -50%
(j) Initial Errors = 807,099 Final Errors = 476,941 -41%
(k) Initial Errors = 15,120 Final Errors = 11,105 -27%
(l) Initial Errors = 358,961 Final Errors = 268,331 -25%
(m) Initial Errors = 383,124 Final Errors = 299,848 -22%
(n) Initial Errors = 9,633,632 Final Errors = 7,808,340 -19%
(o) Initial Errors = 70,254 Final Errors = 58,073 -17%
(p) Initial Errors = 81,421 Final Errors = 68,016 -16%
(q) Initial Errors = 1,125,998 Final Errors = 1,051,712 -7%
(r) Initial Errors = 18,448 Final Errors = 17,686 -4%
(s) Initial Errors = 95,884 Final Errors = 95,877 -0.007%
(t) Initial Errors = 210,170 Final Errors = 210,164 -0.003%
Error Category
Completeness
Conformance
Plausibility

Fig. A1. Each panel tracks the number of errors detected in each iteration of uploading and checking a dataset for a single users' response to a specific data request. On the final iteration, datasets were transferred to the investigator who requested the data.

Table A1
Harmonist Data Toolkit data quality checks based on data model description in REDCap.

	Error	Message to User	Severity
Conformance	**Missing 1st primary key value**	*fieldName* is a primary key variable in this table and should not be blank	Critical
	Invalid PATIENT ID	PATIENT ID was not found in tblBAS. Every PATIENT in other tables should have a record in tblBAS	Critical
	Missing PROGRAM value in tblBAS	Every PATIENT in tblBAS should be linked with an IeDEA Program	Critical
	Deprecated variable	*fieldName* was deprecated over one year ago and is no longer supported in the IeDEA DES	Error
	Invalid format/data type (date)	Date format must be YYYY-MM-DD	Error
	Invalid format/data type (numeric)	Numeric value required for *numericField*	Error
	Invalid code	Invalid code for *codedField*. See <link> for a list of valid codes.	Error
	Invalid PROGRAM	This patient's *PROGRAM* value is not found in tblPROGRAM	Error
	Conflict between _D and _Y	If a date is recorded for xxx_D then xxx_Y must have a value of 1 (Yes)	Error
	Conflict between _RS and Date	If a reason is recorded for a reason (xxx_RS or xxxSTART_RS), then the corresponding date (xxx_D/ xxx_ED or xxx_SD) should not be blank.	Error
	Recently deprecated variable	*fieldName* was deprecated on *date*. Please use *replacement* instead.	Warning
	Deprecated code	*codeValue* is no longer a valid code for *codedField*. It was replaced by *new code*	Warning
	Conflict between date and date approximation	A date was provided but the corresponding date approximation (_A) was coded as Unknown	Warning
Plausibility	**Duplicate PATIENT in single ID table**	PATIENT is the only identifier in this table; every PATIENT value should be unique	Critical
	Duplicate Records	This has key identifier values that duplicate another record in this table.	Error
	Date in future	This date is in the future (Exception: NEXT_VIS_D)	Error
	Date logic error	*Second date* should not be before *first date* [Example: ART_ED before ART_SD, VIS_D before BIRTH_D, or DEATH_D before ENROL_D]	Error
	Value outside expected range	This value is above/below the expected range. The maximum(minimum) value expected for this variable is *upperLimit(lowerLimit)*	Warning
	Conflicting height data	Height should not decrease over time. This patient's height was *height1* on *date1* and *height2* on *date2*.	Warning
Completeness	**Required variable value missing**	*fieldName* is a required variable in this table and should not be blank	Error
	Missing units for lab test	Valid units must be provided for each for *lab value*	Warning
	Missing requested table or column	The table/column requested by *concept xx* is missing from this dataset	Warning
	Conflict between RECART_Y and tblART	This patient is listed in tblBAS as recieving ART (RECART_Y=1) but has no records in tblART	Warning

References

[1] International epidemiology Databases to Evaluate AIDS. https://www.iedea.org/ (accessed March 2, 2021).

[2] V. Huser, F.J. DeFalco, M. Schuemie, P.B. Ryan, N. Shang, M. Velez, R.W. Park, R. D. Boyce, J. Duke, R. Khare, L. Utidjian, C. Bailey, Multisite Evaluation of a Data Quality Tool for Patient-Level Clinical Datasets, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes). 4 (2016) 24. https://doi.org/10.13063/2327-9214.1239.

[3] W.R. Hersh, J. Cimino, P.R.O. Payne, P. Embi, J. Logan, M. Weiner, E.V. Bernstam, H. Lehmann, G. Hripcsak, T. Hartzog, J. Saltz, Recommendations for the use of operational electronic health record data in comparative effectiveness research, EGEMS (Washington, DC) 1 (2013) 1018. https://doi.org/10.13063/2327-9214.1018.

[4] M.G. Kahn, T.J. Callahan, J. Barnard, A.E. Bauck, J. Brown, B.N. Davidson, H. Estiri, C. Goerg, E. Holve, S.G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.C. Ong, P. Ryan, N. Shang, N.G. Weiskopf, C. Weng, M.N. Zozus, L. Schilling, A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data., EGEMS (Washington, DC). 4 (2016) 1244. https://doi.org/10.13063/2327-9214.1244.

[5] T.J. Callahan, A.E. Bauck, D. Bertoch, J. Brown, R. Khare, P.B. Ryan, J. Staab, M. N. Zozus, M.G. Kahn, A Comparison of Data Quality Assessment Checks in Six Data

Sharing Networks, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes) 5 (2017) 8, https://doi.org/10.5334/egems.223.
[6] M.G. Kahn, J.S. Brown, A.T. Chun, B.N. Davidson, D. Meeker, P.B. Ryan, L. M. Schilling, N.G. Weiskopf, A.E. Williams, M.N. Zozus, Transparent Reporting of Data Quality in Distributed Data Networks, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes) 3 (2015) 7. https://doi.org/10.13063/2327-9214.1052.
[7] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J.G. Conde, Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support, J. Biomed. Inform. 42 (2) (2009) 377–381, https://doi.org/10.1016/j.jbi.2008.08.010.
[8] E. Arrivé, S. Ayaya, M.-A. Davies, C. Chimbetete, A. Edmonds, P. Lelo, S.M. Fong, K.A. Razali, K. Kouakou, S.N. Duda, V. Leroy, R.C. Vreeman, Models of support for disclosure of HIV status to HIV-infected children and adolescents in resource-limited settings, J. Int. AIDS Soc. 21 (7) (2018) e25157, https://doi.org/10.1002/jia2.25157.
[9] J. del Amo, All-cause mortality after antiretroviral therapy initiation in HIV-positive women from Europe, Sub-Saharan Africa and the Americas, AIDS. 34 (2020) 277–289, https://doi.org/10.1097/QAD.0000000000002399.
[10] S. Desmonde, F. Tanser, R. Vreeman, E. Takassi, A. Edmonds, P. Lumbiganon, J. Pinto, K. Malateste, C. McGowan, A. Kariminia, M. Yotebieng, F. Dicko, C. Yiannoutsos, M. Mubiana-Mbewe, K. Wools-Kaloustian, M.-A. Davies, V. Leroy, L.M. Mofenson, Access to antiretroviral therapy in HIV-infected children aged 0–19 years in the International Epidemiology Databases to Evaluate AIDS (IeDEA) Global Cohort Consortium, 2004–2015: A prospective cohort study, PLoS Med. 15 (5) (2018) e1002565, https://doi.org/10.1371/journal.pmed.1002565.
[11] E. Zaniewski, O. Tymejczyk, A. Kariminia, S. Desmonde, V. Leroy, N. Ford, A. H. Sohn, D. Nash, M. Yotebieng, M. Cornell, K.N. Althoff, P.F. Rebeiro, M. Egger, IeDEA WHO research-policy collaboration: Contributing real-world evidence to HIV progress reporting and guideline development, J. Virus Erad. 4 (2018) 9–15, https://doi.org/10.1016/S2055-6640(20)30348-4.
[12] O. Tymejczyk, E. Brazier, K. Wools-Kaloustian, M.A. Davies, M. DiLorenzo, A. Edmonds, R. Vreeman, C. Bolton, C. Twizere, N. Okoko, S. Phiri, G. Nakigozi, P. Lelo, P. von Groote, A.H. Sohn, D. Nash, Impact of universal antiretroviral treatment eligibility on rapid treatment initiation among young adolescents with human immunodeficiency virus in Sub-Saharan Africa, J. Infect. Dis. 222 (2020) 755–764, https://doi.org/10.1093/infdis/jiz547.
[13] S.N. Duda, B.S. Musick, M.A. Davies, A.H. Sohn, B. Ledergerber, K. Wools-Kaloustian, C.C. McGowan, N.J. Maxwell, A. Kariminia, C.H.D. Ostinelli, B.C. Hogan, Q. Shi, K. Malateste, R.L. Goodall, D.K. Kristensen, E. V. Hansen, C.F.M. Williams, J.T. Lewis, C.T. Yiannoutsos, The IeDEA data exchange standard: A common data model for global HIV cohort collaboration, MedRxiv. (2020) 2020.07.22.20159921. https://doi.org/10.1101/2020.07.22.20159921.
[14] S.-T. Liaw, J.G.N. Guo, S. Ansari, J. Jonnagaddala, M.A. Godinho, A.J. Borelli, S. de Lusignan, D. Capurro, H. Liyanage, N. Bhattal, V. Bennett, J. Chan, M.G. Kahn, Quality assessment of real-world data repositories across the data life cycle: A literature review, J. Am. Med. Informatics Assoc. 28 (7) (2021) 1591–1599.
[15] L.G. Qualls, T.A. Phillips, B.G. Hammill, J. Topping, D.M. Louzao, J.S. Brown, L. H. Curtis, K. Marsolo, Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet®), EGEMs (Generating Evid. Methods to Improv. Patient Outcomes) 6 (1) (2018) 3, https://doi.org/10.5334/egems.199.
[16] J. Bian, T. Lyu, A. Loiacono, T.M. Viramontes, G. Lipori, Y.i. Guo, Y. Wu, M. Prosperi, T.J. George, C.A. Harle, E.A. Shenkman, W. Hogan, Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data, J. Am. Med. Informatics Assoc. 27 (12) (2020) 1999–2010, https://doi.org/10.1093/jamia/ocaa245.
[17] C.B. Forrest, K.M. McTigue, A.F. Hernandez, L.W. Cohen, H. Cruz, K. Haynes, R. Kaushal, A.N. Kho, K.A. Marsolo, V.P. Nair, R. Platt, J.E. Puro, R.L. Rothman, E.A. Shenkman, L.R. Waitman, N.A. Williams, T.W. Carton, PCORnet® 2020: current state, accomplishments, and future directions, J. Clin. Epidemiol. 129 (2021) 60–67. https://doi.org/10.1016/j.jclinepi.2020.09.036.
[18] Browse Quality Assurance / qa_package - Sentinel Version Control System. https://dev.sentinelsystem.org/projects/QA/repos/qa_package/browse (accessed April 7, 2021).
[19] T.J. Callahan, J.G. Barnard, L.J. Helmkamp, J.A. Maertens, M.G. Kahn, Reporting Data Quality Assessment Results: Identifying Individual and Organizational Barriers and Solutions, EGEMs (Generating Evid. Methods to Improv. Patient Outcomes) 5 (2017) 16, https://doi.org/10.5334/egems.214.
[20] R. Ball, M. Robb, S.A. Anderson, G. Dal Pan, The FDA's sentinel initiative-A comprehensive approach to medical product surveillance, Clin. Pharmacol. Ther. 99 (3) (2016) 265–268, https://doi.org/10.1002/cpt.320.
[21] V. Huser, M.G. Kahn, J.S. Brown, R. Gouripeddi, Methods for examining data quality in healthcare integrated data repositories, in: Pacific Symp. Biocomput., 2018: pp. 628–633.
[22] V. Huser, X. Li, Z. Zhang, S. Jung, R.W. Park, J. Banda, H. Razzaghi, A. Londhe, K. Natarajan, Extending Achilles Heel Data Quality Tool with New Rules Informed by Multi-Site Data Quality Comparison, (2019). https://doi.org/10.3233/SHTI190498.
[23] Creates Descriptive Statistics Summary for an Entire OMOP CDM Instance • Achilles. https://ohdsi.github.io/Achilles/index.html (accessed April 7, 2021).
[24] OHDSI/Achilles: Automated Characterization of Health Information at Large-scale Longitudinal Evidence Systems (ACHILLES) - descriptive statistics about a OMOP CDM database. https://github.com/OHDSI/Achilles (accessed March 3, 2021).
[25] C. Blacketer, F.J. Defalco, P.B. Ryan, P.R. Rijnbeek, Increasing Trust in Real-World Evidence Through Evaluation of Observational Data Quality, MedRxiv. (2021) 2021.03.25.21254341. https://doi.org/10.1101/2021.03.25.21254341.
[26] R Core Team (R Foundation for Statistical Computing), R: A language and environment for statistical computing. https://www.r-project.org/ (accessed March 3, 2021).
[27] R. Khare, L. Utidjian, B.J. Ruth, M.G. Kahn, E. Burrows, K. Marsolo, N. Patibandla, H. Razzaghi, R. Colvin, D. Ranade, M. Kitzmiller, D. Eckrich, L.C. Bailey, A longitudinal analysis of data quality in a large pediatric data research network, J. Am. Med. Informatics Assoc. 24 (2017) 1072–1079, https://doi.org/10.1093/jamia/ocx033.
[28] et al. Chang W, Cheng J, Allaire J, shiny: Web Application Framework for R [R package version 1.6.0], (2021).
[29] J. Allaire, Y. Xie, J. McPherson, rmarkdown: Dynamic Documents for R. R package version 2.7. https://rmarkdown.rstudio.com (accessed May 3, 2021).
[30] H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York (2016), https://doi.org/10.1007/978-0-387-98141-3.
[31] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the Tidyverse, J. Open Source Softw. 4 (2019) 1686. https://doi.org/10.21105/joss.01686.
[32] HL7 Fast Heathcare Interoperability Resources v4.0.1. https://www.hl7.org/fhir/index.html (accessed January 24, 2022).
[33] NCBO BioPortal. https://bioportal.bioontology.org/ (accessed November 3, 2021).
[34] IeDEA/Harmonist: IeDEA Harmonist Data Toolkit. https://github.com/IeDEA/Harmonist (accessed March 3, 2021).